**Лабораторная работа 2–3. Анализ данных с помощью операций трансформации**

**Цель работы** — научиться выполнять анализ данных в R с помощью операций трансформации из пакета dplyr: filter, arrange, select, mutate, summarise, group\_by.

**Вспомогательный материал:** Лекция 4.

**Дополнительные файлы:** https://goo.gl/1N7ePq

**Общие указания:**

1. В качестве отчета о выполнении практического занятия студент представляет преподавателю программный код (***lr2-3.R***).
2. Для импорта, экспорта и обработки данных нельзя использовать возможности интерфейса RStudio — можно только писать программный код.
3. Задания должны быть выполнены наиболее оптимальным образом (автоматизация, минимизация строк кода, универсальность и др.).
4. Программный код должен быть оформлен в соответствии с Google's R Style Guide.
5. Смысловые блоки программного кода необходимо сопровождать комментариями.

**Задание 1**

***—> transform\_lr2.R***

Откройте скрипт transform\_lr2.R. Скопируйте код в свой файл.

Ознакомьтесь с кодом в разделе **filter**.

Найдите все рейсы, для которых (которые):

* Время задержки прибытия (arrival delay) равно или превышает два часа.
* Отправлялись в Houston (IAH или HOU).
* Находятся в ведении операторов UnitedAirlines, American Airlines, or Delta Airlines.
* Отправлялись летом.
* Прибыли позже более чем на 2 часа, но отправлялись вовремя.
* Задержались, по крайней мере, на час, но наверстали более 30 мнут в течение полета.
* Отправлялись между полночью и 6 часами утра (включительно).

В каких случаях можно применить функцию between()?

Для скольких рейсов отсутствует значение dep\_time?

Для каких ещё переменных имеются отсутствующие значения?

**Задание 2**

Ознакомьтесь с кодом в разделе **arrange**.

Отсортируйте данные:

* По какой-либо переменной так, чтобы все NA были вначале (is.na()).
* Чтобы вначале оказались рейсы с наибольшим временем задержки.
* Чтобы вначале оказались рейсы, которые отправлялись наиболее раньше планируемого времени.
* Чтобы вначале оказались наименее длительные рейсы.

**Задание 3**

Ознакомьтесь с кодом в разделе **select**.

Реализуйте как можно больше способов выбрать dep\_time, dep\_delay, arr\_time и arr\_delay.

Что будет, если вызвать несколько раз переменную в select()?

Чем функция one\_of() может быть полезна в сочетании с вектором vars <- c("year", "month", "day", "dep\_delay", "arr\_delay")?

**Задание 4**

Ознакомьтесь с кодом в разделе **mutate**.

Преобразуйте данные в столбцах dep\_time, sched\_dep\_time, arr\_time, sched\_arr\_time, air\_time (создайте новые переменные) так, чтобы время отображалось в привычном виде (сейчас 845 означает 8 часов 45 минут).

Сравните air\_time и arr\_time - dep\_time.

Сравните dep\_time, sched\_dep\_time и dep\_delay.

Найдите 10 рейсов с наибольшей задержкой (используйте min\_rank()).

**Задание 5**

Ознакомьтесь с кодом в разделе **summarise**.

Оцените характеристики задержки по пяти различным группам рейсов:

1. Рейс вылетает на 15 минут раньше в 50% случаев и прилетает на 15 минут позже в 50% случаев.
2. Прибытие рейса всегда задерживается на 10 минут.
3. В 99% случаев рейс совершается вовремя. В 1% случаев задерживается на 2 часа.
4. Рейсы какого оператора задерживаются больше всего?
5. Для каждого самолета посчитайте количество рейсов до первого случая более чем часовой задержки.

*Литература:*

*Grolemund, G. R for Data Science [Electronic resource] / Garrett Grolemund, Hadley Wickham. – 2016. – Mode of access: http://r4ds.had.co.nz/index.html. – Date of access: 01.09.2016.*